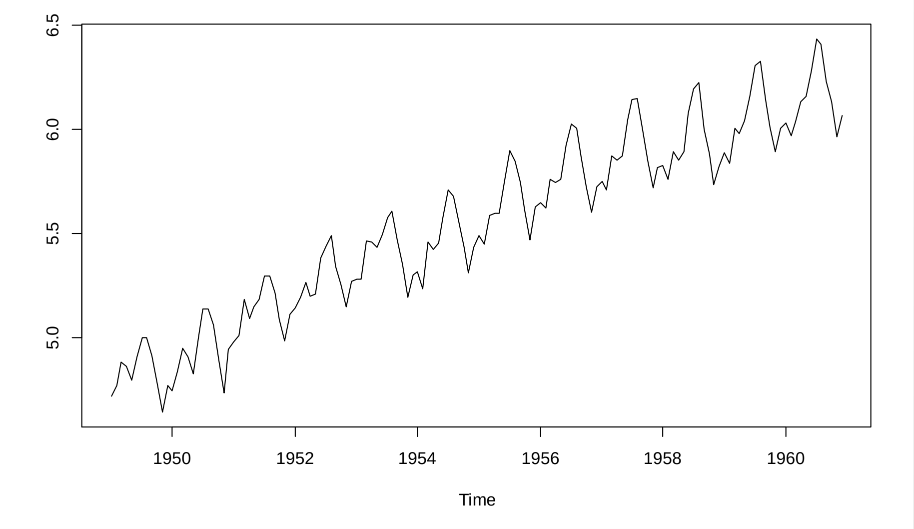
제 2장 추세분석

회귀모형에 의한 시계열분석

회귀모형에 의한 시계열분석

- 추세 및 계절변동이 비교적 규칙적인 경우



\- 추세와 계절변동을 회귀모형으로 설명

일반적인 회귀분석과의 차이점

- 일반적인 회귀모형, 즉 OLS(Ordinary Least Squares) 회귀모형에서는 오차가 서로 독립임을 가정
- 시계열자료의 특성으로 서로 독립인 오차를 가정하는 것은 불가능
- 오차 사이의 상관관계, 즉 자기상관을 설명하기 위한 추후조치가 필요

시계열 자료에 대한 회귀모형

- 추세 성분만이 있는 모형: 다항회귀모형

$$Z_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + \varepsilon_t$$

- 계절 성분만이 있는 모형: 지시변수에 의한 회귀모형

$$Z_t = \sum_{i=1}^{s} \beta_i D_{t,i} + \varepsilon_t, \qquad D_{t,i} = \begin{cases} 1, & t = i (\mathrm{mod}\ s) \\ 0, & \text{otherwise} \end{cases}$$

- 계절 성분 주기: s
- 월별 자료($s = 12$)에 12개 지시변수 사용
- β_i: i 월의 효과(평균)

t	년 월	$D_{t,1}$	$D_{t,2}$	$D_{t,3}$	. . .	$D_{t,12}$
1	2012.01	1	0	0	0	0
2	2012.02	0	1	0	0	0
3	2012.03	0	0	1	0	0
:	:	:	:	:	:	:
12	1212.12	0	0	0	0	1
13	1213.01	1	0	0	0	0
14	1213.02	0	1	0	0	0

- 1차 추세와 계절 성분이 함께 있는 모형:

$$Z_t = \beta_1 t + \sum_{i=2}^{s+1} \beta_i D_{t,i} + \varepsilon_t, \qquad D_{t,i} = \begin{cases} 1, & t = i (\mathrm{mod}\ s) \\ 0, & \text{otherwise} \end{cases}$$

  - 절편 β_0 제거된 모형. β_i: i 월의 효과(평균)
  - 절편을 포함하려면, 지시변수를 $(s-1)$개 사용해야 함. 이 경우 β_i는 다른 의미로 해석됨.

- 시계열 자료와 같이 오차가 서로 독립이 아닌 경우의 모형:

$$Z_t = \beta_1 t + \sum_{i=2}^{s+1} \beta_i D_{t,i} + \varepsilon_t, \qquad \varepsilon_t \sim ARMA(p,q)$$

$$\phi(B)\varepsilon_t = \theta(B) w_t, \quad w_t \sim WN(0, \sigma^2)$$

2.1 회귀모형

2.1.1 회귀모형

2.1.2 최소제곱에 의한 모수추정

2.1.3 최소제곱추정량의 성질들

2.1. 4 구간추정 및 가설검정

- 내용은 교재 참조(생략)

2.1.5 잔차분석

- 일반적인 회귀모형에서 오차 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 에 대한 가정

$$\varepsilon_i \quad iid \quad N(0, \sigma^2)$$

- 가정만족 여부 확인
  - 잔차 시계열그림(교재 46쪽 그림 2-2)
  - 잔차 QQ-plot (혹은 히스토그램)
  - 잔차
    - ACF
    - portmanteau test(Ljung-Box test)
    - Durbin-Watson test

Durbin-Watson(DW) 검정

- 오차항의 1차 자기상관 존재 여부에 대한 통계적 검정
- 귀무가설 $H_0 : \rho_1 = 0$
- 검정통계량

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \cong 2(1 - \hat{\rho}_1)$$

- p값 계산 방법:

  패키지 car의 함수 durbinWatsonTest( )

- Ljung-Box test가 더 포괄적 검정

2.2 다항추세 + ARMA 오차 회귀모형

예 2-1: 선형추세모형의 적합 예제

- 데이터 파일: pop.txt
- 내용: 1960~1995 우리나라 총인구(연도별 자료)
- 만 명 단위로 분석

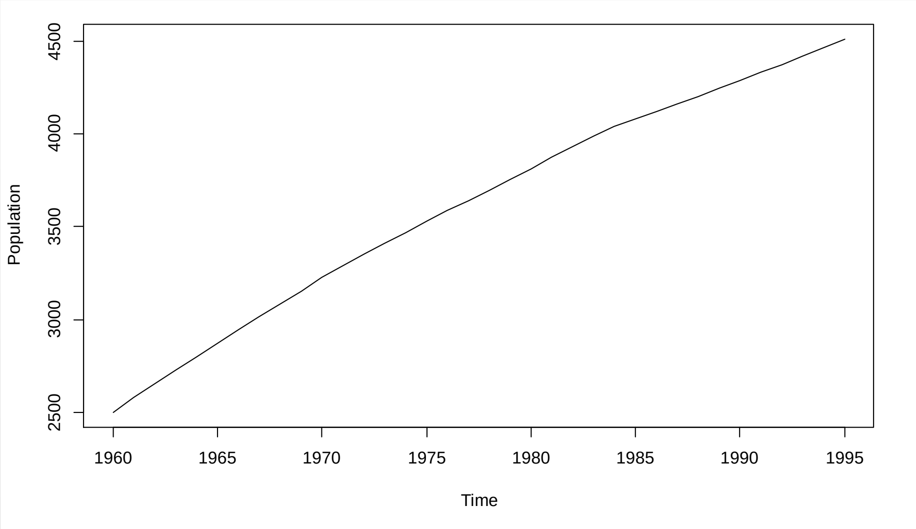
- 선형추세회귀모형 + ARMA 오차 모형

추세모형: $Z_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + \varepsilon_t$

오차 모형: $\phi(B)\varepsilon_t = \theta(B) w_t, \quad w_t \sim WN(0, \sigma^2)$

- 자료 입력 및 시계열 그림 작성

```
> pop <- scan("D:/Data/pop.txt")
> pop <- round(pop/10000)
> pop.ts <- ts(pop , start = 1960, freq = 1)
> plot(pop.ts , ylab="Population")
```



- 1차 추세 적합 시도

$$Z_t = \beta_0 + \beta_1 t + \varepsilon_t$$

- 변수 t의 생성 및 회귀모형 적합

```
> Time <- time(pop.ts)
> fit1 <- lm(pop.ts~Time)
```

- 적합 결과 확인

```
> summary(fit1)

Residuals:
    Min      1Q  Median      3Q     Max 
-115.40  -48.30   16.87   54.37   63.29 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.091e+05  1.868e+03  -58.43   <2e-16 ***
Time         5.701e+01  9.444e-01   60.37   <2e-16 ***

Residual standard error: 58.87 on 34 degrees of freedom
Multiple R-squared:  0.9908,	Adjusted R-squared:  0.9905 
F-statistic:  3644 on 1 and 34 DF,  p-value: < 2.2e-16
```

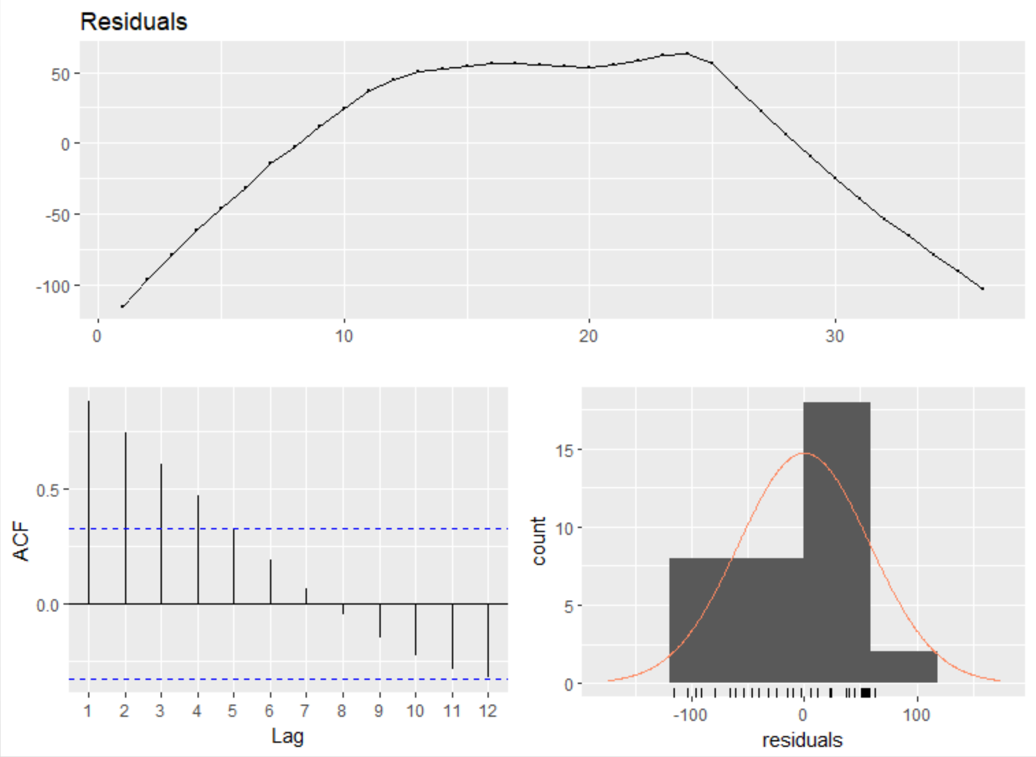
- 교재에서 사용된 변수 t 생성
  Time <- 1:length(pop.ts)

- 교재의 절편 추정 값과 다른 이유

- 잔차 분석

```
> library(forecast)
> checkresiduals(fit1)
```

```
        Breusch-Godfrey test for serial correlation of order up to 10
data:  Residuals
LM test = 31.138, df = 10, p-value = 0.0005568
```



- 2차 추세?
- 양의 상관관계?

- 2차 추세모형 시도

- 2차 추세모형의 적합

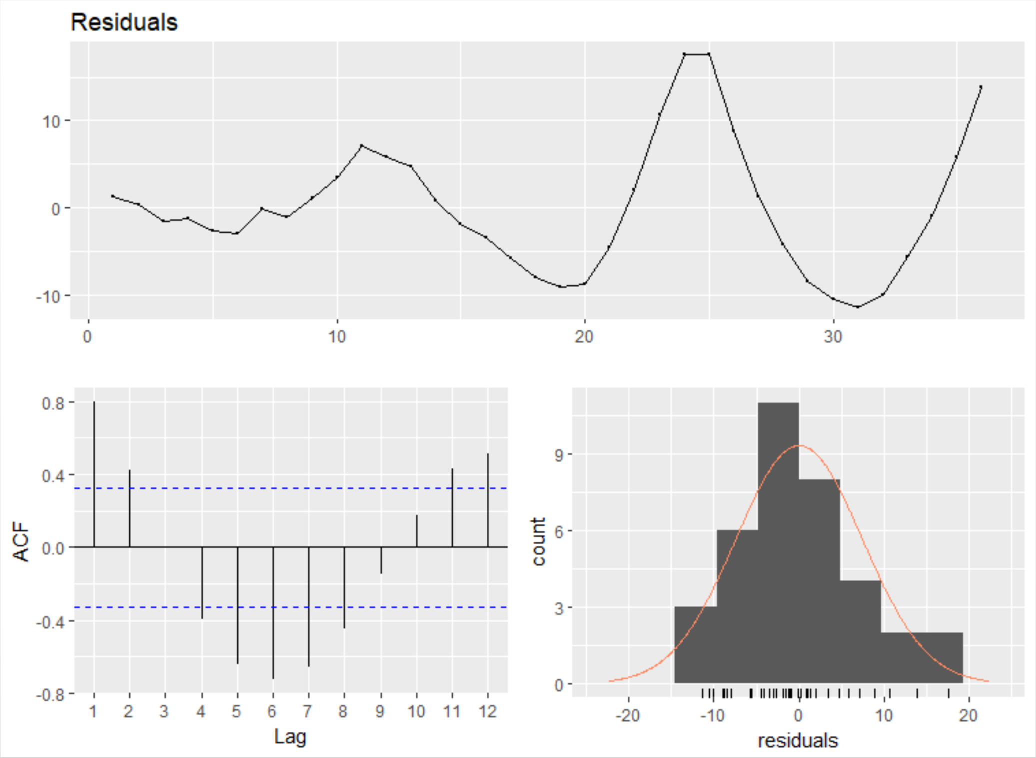
```
> fit2 <- lm(pop.ts~Time+I(Time^2))
```

- R 공식에서 함수 I()의 역할:
  `lm(y~x1+x2)`
  `lm(y~I(x1+x2))`

```
> summary(fit2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.410e+06  5.185e+04  -46.49   <2e-16 ***
Time         2.384e+03  5.244e+01   45.47   <2e-16 ***
I(Time^2)   -5.885e-01  1.326e-02  -44.38   <2e-16 ***

Residual standard error: 7.67 on 33 degrees of freedom
Multiple R-squared:  0.9998,	Adjusted R-squared:  0.9998
F-statistic: 1.083e+05 on 2 and 33 DF,  p-value: < 2.2e-16
```

- 2차 추세모형의 잔차분석

```
> checkresiduals(fit2)

data:  Residuals
LM test = 34.622, df = 10, p-value = 0.0001448
```



- 분산 증가
- 로그 변환 필요

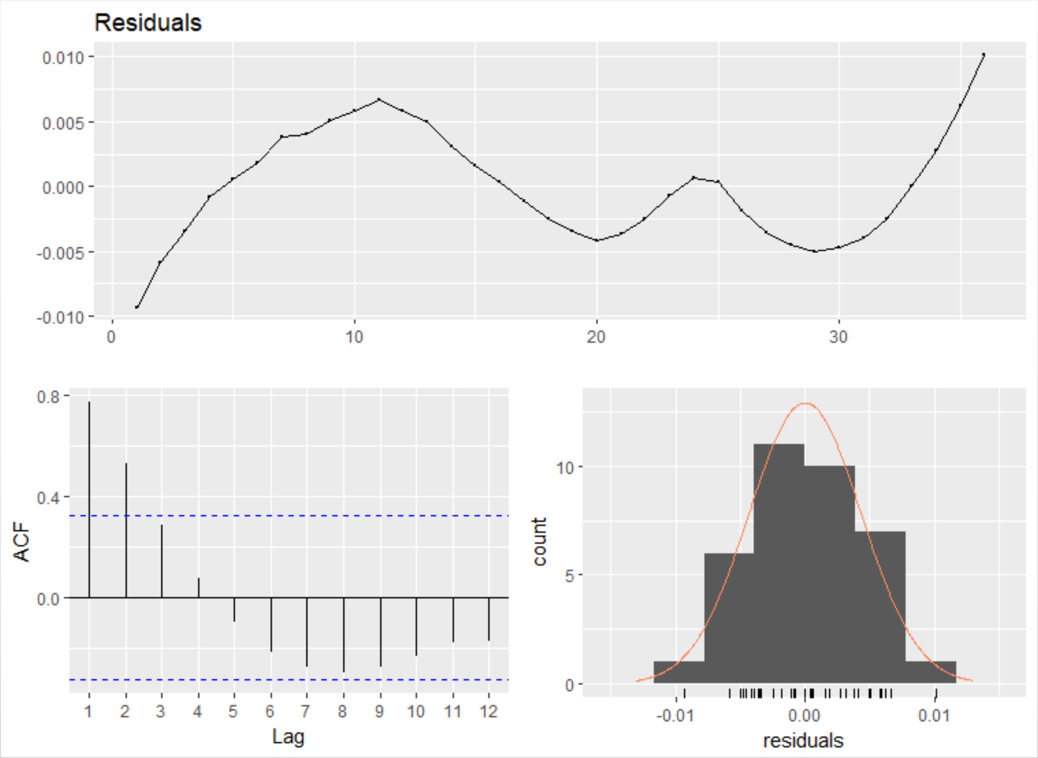
- 로그 변환 후 2차 추세모형의 적합

```
> fit3 <- lm(log(pop.ts)~Time+I(Time^2))
> summary(fit3)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.199e+03  3.016e+01  -39.75   <2e-16 ***
Time         1.204e+00  3.050e-02   39.49   <2e-16 ***
I(Time^2)   -3.004e-04  7.712e-06  -38.95   <2e-16 ***
---
Residual standard error: 0.004461 on 33 degrees of freedom
Multiple R-squared:  0.9994,	Adjusted R-squared:  0.9993 
F-statistic: 2.664e+04 on 2 and 33 DF,  p-value: < 2.2e-16
```

- 잔차 분석

```
> checkresiduals(fit3)

data:  Residuals
LM test = 28.04, df = 10, p-value = 0.001779
```



- 잔차: 양의 자기 상관
- 일정기간 양의 값 &
  일정기간 음의 값

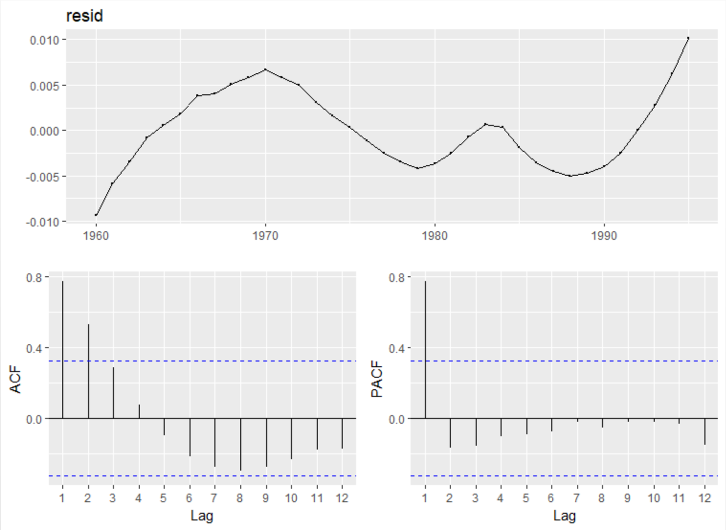
- 오차가 독립이 아님
- 오차에 대한 모형이 필요함

- 오차의 모형: ARMA(p,q)

- 오차 모형 단계
  1) 모형 식별
  2) 모수 추정
  3) 모형 진단

- 오차의 모형 식별

```
> resid <- ts(fit3$resid,start=1960)
> ggtsdisplay(resid)
```



AR(1) 식별

• 추정 및 진단

```
> fit_r1 <- Arima(resid,order=c(1,0,0),include.mean=FALSE)
> checkresiduals(fit_r1)

data:  Residuals from ARIMA(1,0,0) with zero mean
Q* = 34.785, df = 9, p-value = 6.503e-05
```

```
> fit_r2 <- Arima(resid,order=c(2,0,0),include.mean=FALSE)
> fit_r3 <- Arima(resid,order=c(1,0,1),include.mean=FALSE)

> confint(fit_r2)
        2.5 %     97.5 %
ar1  1.770756  1.9740120
ar2 -1.038370 -0.8609524

> confint(fit_r3)
        2.5 %   97.5 %
ar1 0.8861058 1.049141
ma1 0.5594046 0.910749
```

- 과대 적합 모형의 진단

```
> checkresiduals(fit_r2)

data:  Residuals from ARIMA(2,0,0) with zero mean
Q* = 8.8702, df = 8, p-value = 0.3534

> checkresiduals(fit_r3)

data:  Residuals from ARIMA(1,0,1) with zero mean
Q* = 20.604, df = 8, p-value = 0.008278
```

AR(2): 가정 만족

오차모형: AR(2)

- AR(2) 모형의 과대적합

```
> confint(Arima(resid,order=c(2,0,1),include.mean=FALSE))
          2.5 %     97.5 %
ar1  1.7436707  1.9825692
ar2 -1.0500424 -0.8288453
ma1 -0.2178853  0.4054730
> confint(Arima(resid,order=c(3,0,0),include.mean=FALSE))
          2.5 %     97.5 %
ar1  1.6411158  2.3518888
ar2 -1.8677351 -0.5185865
ar3 -0.2308685  0.4937741
```

- 추세모형(fit3) + AR(2) 오차모형: 두 모형의 결합

```
> fit_x <- model.matrix(fit3)[,-1]
> f1 <- Arima(pop.ts,order=c(2,0,0),xreg=fit_x,
              lambda=0)
```

```
> f1
Series: pop.ts
Regression with ARIMA(2,0,0) errors
Box Cox transformation: lambda= 0

Coefficients:
         ar1      ar2   intercept    Time   Time2
      1.8665  -0.9234  -1183.5728  1.1886  -3e-04
s.e.  0.0611   0.0581      8.2792  0.0087   0e+00

sigma^2 estimated as 6.179e-07:  log likelihood=205.62
AIC=-399.24   AICc=-396.34   BIC=-389.73
```

추세모형:
$Y = X\beta + \varepsilon$

함수 Arima( )
'xreg='에 행렬 X 에서 1로 이루어진 첫 번째 열을 제외한 행렬을 지정

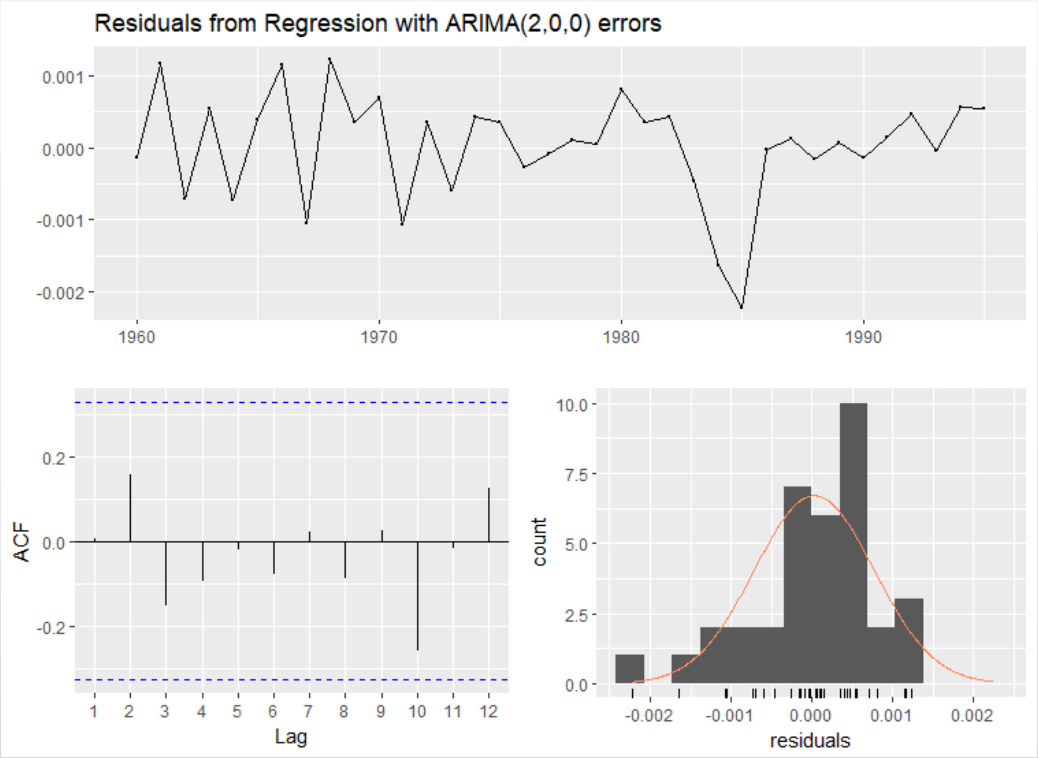
행렬 X의 추출
model.matrix( )에 추세모형 입력

모형식: $$\log(Z_t) = -1183.6 + 1.19t - 0.0003t^2 + e_t$$

$$e_t = 1.87e_{t-1} - 0.93e_{t-2}$$

- 최종 모형의 잔차 분석

```
> checkresiduals(f1)

data:  Residuals from Regression with ARIMA(2,0,0) errors
Q* = 6.6009, df = 5, p-value = 0.2521
```

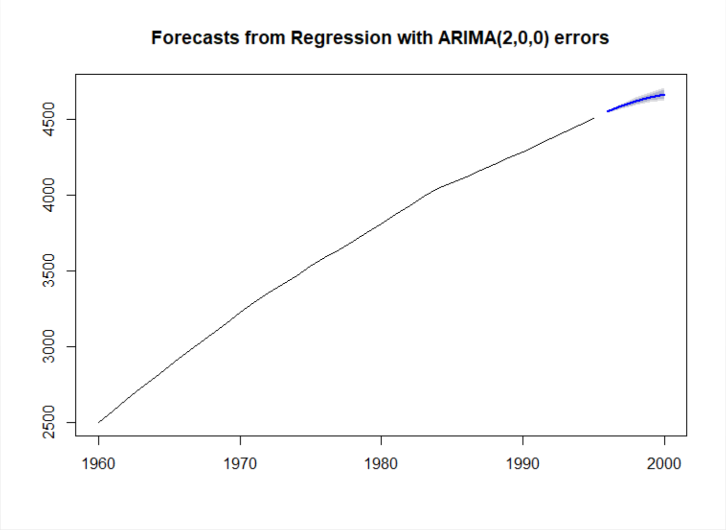


- 백색잡음과정 확인
- 정규분포

- 최종 모형의 예측

```
> new_x <- time(ts(start=1996,end=2000))
> fore_1 <- forecast(f1,xreg=cbind(new_x,new_x^2))
```

```
> plot(fore_1)
```



- 'xreg=' : 최종모형에 포함된 X 변수의 행렬
- 예측 기간에 대한 변수 t와 t^2을 행렬로 구성

```
> accuracy(fore_1)
                     ME  RMSE   MAE      MPE    MAPE    MASE   ACF1
Training set 0.05364 2.621 1.902 0.002791 0.05456 0.03316 0.1606
```

2.2.4 계절추세 + ARMA 오차 회귀모형

$$Z_t = \beta_1 t + \sum_{i=2}^{s+1} \beta_i D_{t,i} + \varepsilon_t, \qquad D_{t,i} = \begin{cases} 1, & t = i (\mathrm{mod}\ s) \\ 0, & \text{otherwise} \end{cases}$$

$$\phi(B)\varepsilon_t = \theta(B) w_t, \quad w_t \sim WN(0, \sigma^2)$$

예제 2-2: 추세와 계절 성분을 동시에 갖는 모형 적합 예제

- 데이터 파일: depart.txt
- 내용: 1984년 1월부터 1988년 12월까지 어떤 백화점의 월별 매출액
- 계절 변동 요인을 지시변수를 사용하여 모형 적합

- 지시변수에 의한 계절추세성분 모형화

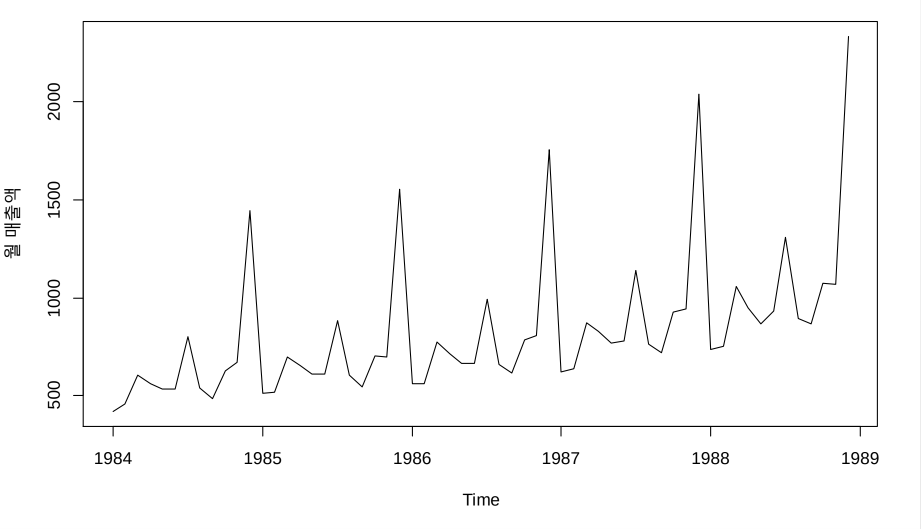
비교: 2.4절 자기회귀오차모형

계절추세모형: $$\log(Z_t) = \beta_1 t + \sum_{i=2}^{13} \beta_i D_{t,i} + \varepsilon_t$$

오차 모형: $$\phi(B)\varepsilon_t = \theta(B) w_t, \quad w_t \sim WN(0, \sigma^2)$$

- 자료 입력 및 시계열 그림 작성

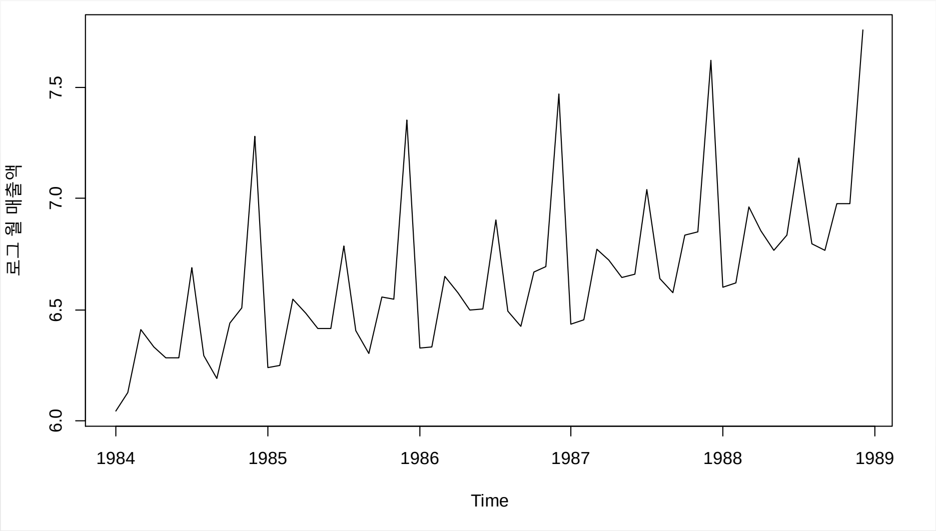
```
> depart <- scan("D:/Data/depart.txt")
> depart.ts <- ts(depart, start=c(1984,1), freq=12)
> plot(depart.ts, ylab="월 매출액")
```



- 뚜렷한 추세 및 계절성분
- 분산증가

- 분산 안정화 후 시계열 그림 작성

```
> lndepart <- log(depart.ts)
> plot(lndepart, ylab="로그 월 매출액")
```



- 계절추세모형 적합

  - 변수 t와 $D_{t,i}$의 생성

```
> Time <- time(lndepart)
> Month <- cycle(lndepart)
```

  - 계절추세모형 적합

```
> fit1 <- lm(lndepart~Time+factor(Month)+0)
```

factor(Month) : 12개 지시변수 모형에 추가
0 : 절편 제거

- 적합 결과

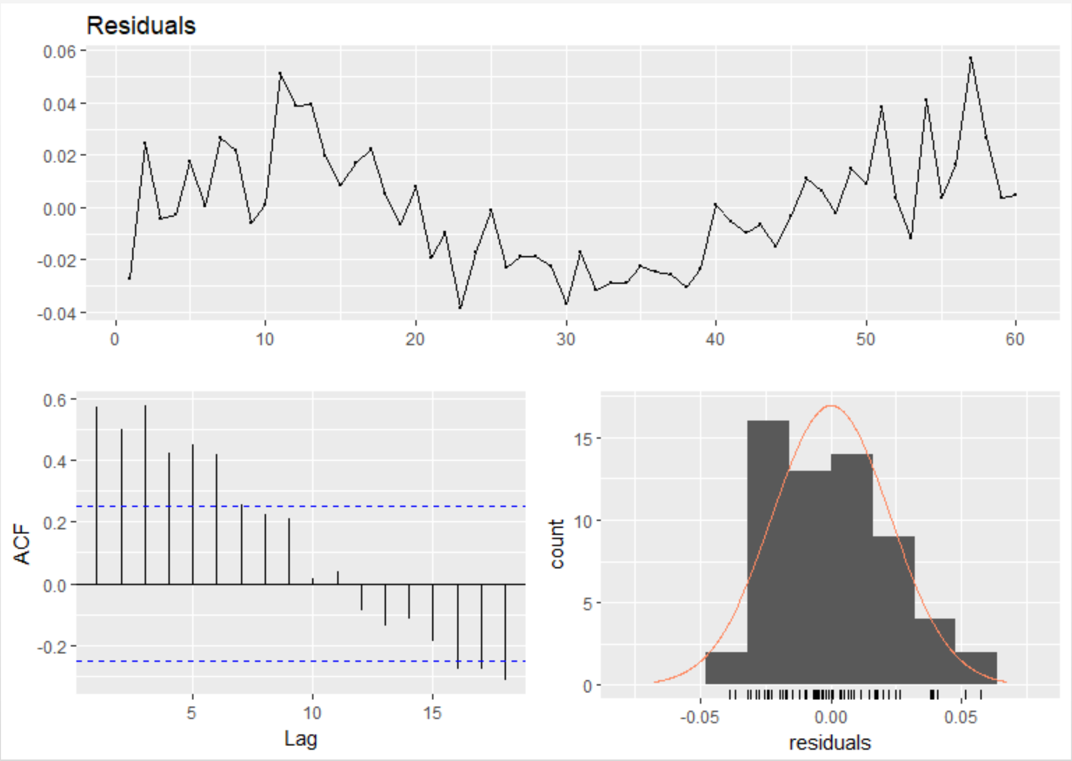
```
> summary(fit1)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
Time                0.12792    0.00231   55.39   <2e-16 ***
factor(Month)1  -247.72500    4.58685  -54.01   <2e-16 ***
factor(Month)2  -247.70839    4.58705  -54.00   <2e-16 ***
factor(Month)3  -247.40807    4.58724  -53.93   <2e-16 ***
factor(Month)4  -247.49384    4.58743  -53.95   <2e-16 ***
factor(Month)5  -247.57595    4.58762  -53.97   <2e-16 ***
factor(Month)6  -247.56941    4.58782  -53.96   <2e-16 ***
factor(Month)7  -247.20068    4.58801  -53.88   <2e-16 ***
factor(Month)8  -247.60491    4.58820  -53.97   <2e-16 ***
factor(Month)9  -247.68907    4.58839  -53.98   <2e-16 ***
factor(Month)10 -247.45574    4.58859  -53.93   <2e-16 ***
factor(Month)11 -247.44748    4.58878  -53.92   <2e-16 ***
factor(Month)12 -246.67871    4.58897  -53.76   <2e-16 ***
---
Residual standard error: 0.0253 on 47 degrees of freedom
Multiple R-squared:      1,       Adjusted R-squared:      1
F-statistic: 3.199e+05 on 13 and 47 DF,  p-value: < 2.2e-16
```

지시변수 중 비유의적인 것이 있어도 제거하면 안됨

- 잔차분석

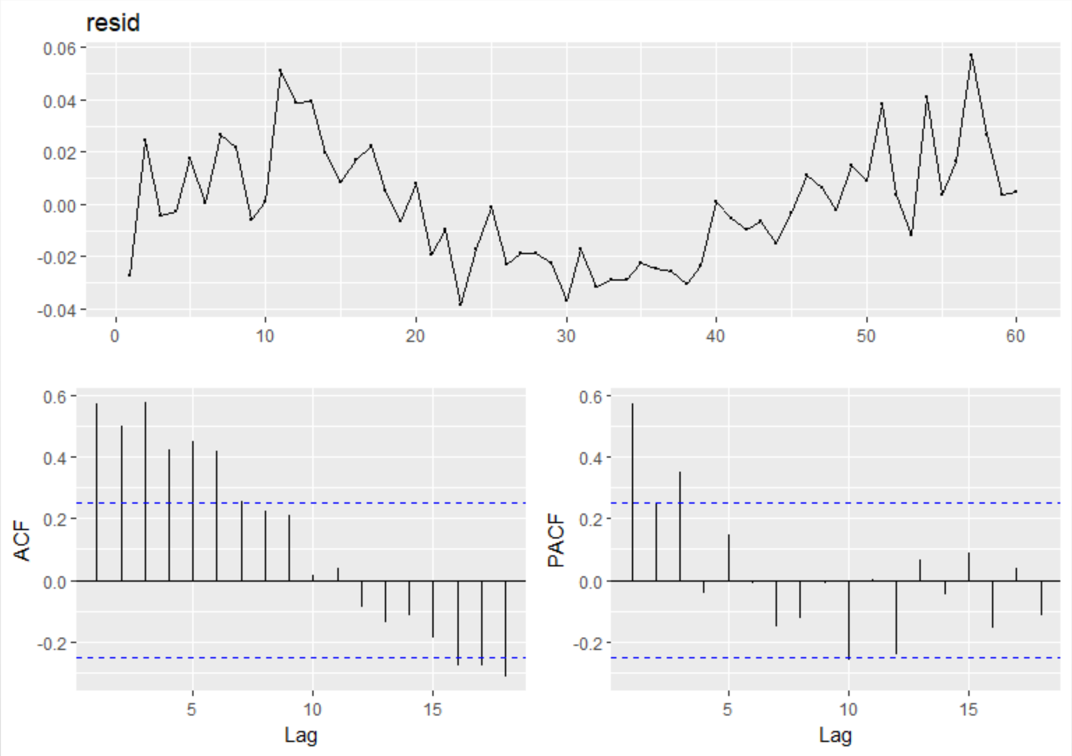
```
> checkresiduals(fit1)

data:  Residuals
LM test = 36.542, df = 16, p-value = 0.002432
```



오차에 대한 모형이 필요

- 오차모형 식별

```
> resid <- fit1$residuals
> ggtsdisplay(resid)
```



AR(3) 식별

- 오차 모형 추정

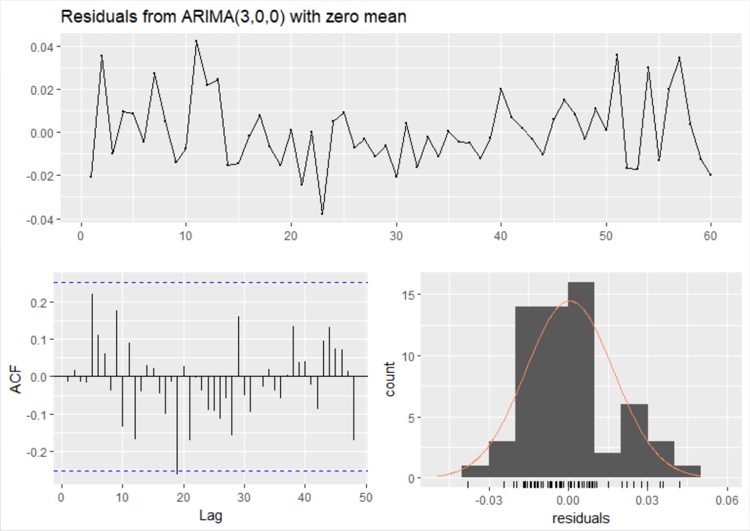
```
> Arima(resid,order=c(3,0,0),include.mean=FALSE)

Coefficients:
         ar1     ar2     ar3
      0.3511  0.0497  0.3822
s.e.  0.1231  0.1371  0.1251

sigma^2 estimated as 0.0002842:  log likelihood=160.91
AIC=-313.82   AICc=-313.1   BIC=-305.45
```

ar2 비유의적

```
> fit_r1 <- Arima(resid,order=c(3,0,0),include.mean=FALSE,
                  fixed=c(NA,0,NA))
> fit_r1

Coefficients:
         ar1  ar2     ar3
      0.3718    0  0.4033
s.e.  0.1093    0  0.1107

sigma^2 estimated as 0.0002848:  log likelihood=160.85
AIC=-315.69   AICc=-315.26   BIC=-309.41
```

- 오차 모형 잔차 분석

```
> checkresiduals(fit_r1,lag.max=48,lag=24)

data:  Residuals from ARIMA(3,0,0) with zero mean
Q* = 22.56, df = 21, p-value = 0.3679

Model df: 3.   Total lags used: 24
```



- AR(3) 모형의 과대적합

```
> confint(Arima(resid,order=c(3,0,1),include.mean=FALSE,
          fixed=c(NA,0,NA,NA)))
          2.5 %    97.5 %
ar1 -0.16881705 0.9289584
ar2          NA        NA
ar3  0.06583327 0.7321927
ma1 -0.70846193 0.6855687

> confint(Arima(resid,order=c(4,0,0),include.mean=FALSE,
                fixed=c(NA,0,NA,NA)))
         2.5 %    97.5 %
ar1  0.1474370 0.6341337
ar2         NA        NA
ar3  0.1787318 0.6634399
ar4 -0.3304956 0.2376827
```

오차 최종 모형: ar2가 비유의적인 AR(3)

- 계절추세모형(fit1)과 AR(3) 오차 모형의 결합: 최종 모형

```
> fit_x <- model.matrix(fit1)
> fit2 <- Arima(depart.ts,order=c(3,0,0),xreg=fit_x,
                include.mean=FALSE,
                fixed=c(NA,0,NA,rep(NA,13)),lambda=0)
```

fit1: 절편이 없는 모형
fit1 + ar2가 비유의적인 AR(3) 모형의 모수: ar1, ar2, ar3, t, D1, … , D12

- 모수 추정 결과 및 유의성 검정

```
> coef(fit2)
            ar1             ar2             ar3            Time
      0.3721191       0.0000000       0.4145015       0.1302350
factor(Month)1  factor(Month)2  factor(Month)3  factor(Month)4
   -252.3139317    -252.2990950    -251.9993784    -252.0846335
factor(Month)5  factor(Month)6  factor(Month)7  factor(Month)8
   -252.1672680    -252.1617338    -251.7928366    -252.1966632
factor(Month)9 factor(Month)10 factor(Month)11 factor(Month)12
   -252.2827161    -252.0496915    -252.0385725    -251.2736958
```

```
> confint(fit2)
```

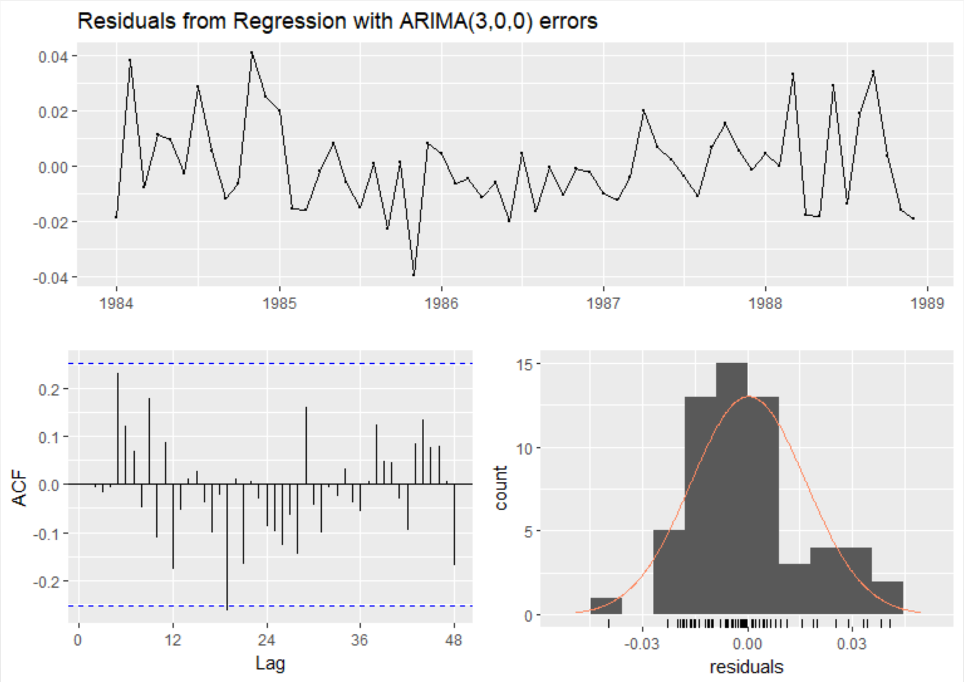
모수 모두 유의적

모형식: $\log(Z_T) = 0.13t - 252.2D_1 - \cdots - 251.3D_{12} + e_t$

$$e_t = 0.37e_{t-1} + 0.41e_{t-3}$$

- 최종 모형의 잔차분석

```
> checkresiduals(fit2,lag.max=48)

data:  Residuals from Regression with ARIMA(3,0,0) errors
Q* = 22.653, df = 8, p-value = 0.00384

Model df: 16.   Total lags used: 24
```



- 검정 결과: 독립 가설 기각
- 이유: 최종 모형에 포함된 변수가 많아서 검정 자유도가 너무 작아서 발생한 것으로 판단

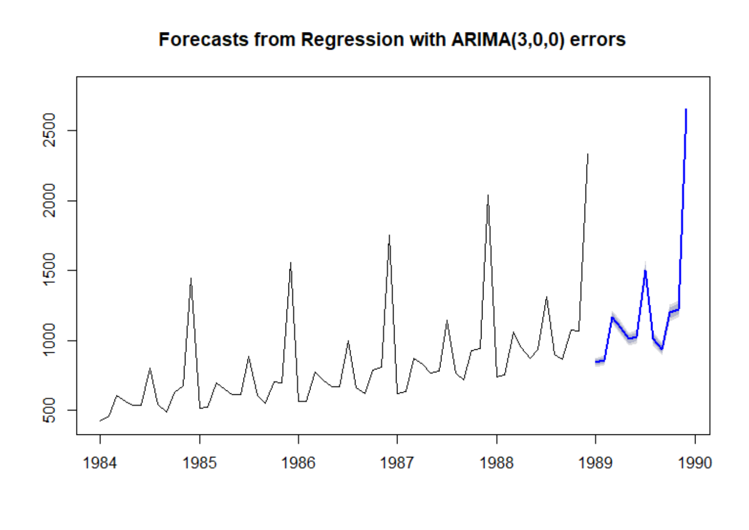
- 잔차의 ACF: 독립성에 큰 문제가 없는 것으로 보임

- 최종모형에 의한 예측

```
> new_t <- time(ts(start=c(1989,1),end=c(1989,12),freq=12))
> new_x <- cbind(new_t,diag(rep(1,12)))
> fore_2 <- forecast(fit2,xreg=new_x)
```

  - 예측 기간에 해당되는 변수 t와 지시변수 $D_1, \cdots, D_{12}$로 행렬 구성

```
> plot(fore_2)
```



- 예측 정확성 측도

```
> accuracy(fore_2)
                    ME     RMSE      MAE         MPE    MAPE      MASE
Training set 0.160595 14.2812 10.3467 0.0154101 1.2583 0.096661
                   ACF1
Training set 0.0625630
```

- 참고: 계절형 ARIMA 최종 모형의 예측 정확성 측도

```
> accuracy(fore3_1)
                  ME   RMSE    MAE    MPE  MAPE   MASE    ACF1
Training set -1.937 16.93 11.60 -0.20 1.36 0.108 0.0435
```